Title: Phonics: Reading policy and the evidence of effectiveness from a systematic 'tertiary' review

Torgerson, C.[a]*, Brooks, G.[b], Gascoine, L.[a] and Higgins, S.[a]

[a]School of Education, Durham University, Durham, DH1 1TA

[b]School of Education, University of Sheffield, Sheffield, S10 2GU

*corresponding author

carole.torgerson@durham.ac.uk

Biographical notes:

Name: Carole Torgerson

Professor of Education, Durham University

Leazes Road, Durham University

0191 334 8382

carole.torgerson@durham.ac.uk

Professor Carole Torgerson has been a Professor of Education at Durham University since 2012. Prior to this she was Professor of Experimental Design at the University of Birmingham and Reader in Evidence-based Education at the University of York. She is an expert on randomised controlled trial and systematic review designs, having undertaken numerous experiments and reviews in various topics in education. She is also a literacy expert.

Name: Greg Brooks

Emeritus Professor of Education, University of Sheffield

4 Peel Terrace, Sheffield S10 2GU

0114 267 0097

g.brooks@sheffield.ac.uk

At NFER (1981-2000) **Professor Greg Brooks** worked on oracy assessment and family literacy evaluations. At Sheffield (2001-07) he directed 15 adult literacy projects. In 2005-06 he was a member of the Rose committee, and in 2008-09 of the dyslexia subgroup of the Rose review of the primary curriculum in England. In 2011-12 he was a member of the EU High Level Group of Experts on Literacy.

Name: Louise Gascoine

Research Associate, Durham University

Leazes Road, Durham

0191 334 8382

louise.gascoine@durham.ac.uk

Dr Louise Gascoine is a Research Associate at Durham University. She has a PhD in education (focused on metacognition). Her current research is focused on metacognition, systematic review design and the use of impact and process evaluations within randomised controlled trial design in education.

Name: Steve Higgins

Professor, Durham University

Leazes Road, Durham

Professor Steve Higgins is a former primary school teacher. His research interests include the effective use of digital technologies for learning in schools, understanding how children's thinking and reasoning develop, and how teachers can be supported in developing the quality and effectiveness of teaching and learning in their classrooms, using evidence from research.

0191 334 8359

s.e.higgins@durham.ac.uk

Phonics: Reading policy and the evidence of effectiveness from a systematic 'tertiary' review

Abstract

Ten years after publication of two reviews of the evidence on phonics (Rose, 2006; Torgerson *et al.*, 2006), a number of British policy initiatives have firmly embedded phonics in the curriculum for early reading development. However, uncertainty about the most effective approaches to teaching reading remains. A definitive trial comparing different approaches was recommended in 2006, but never undertaken. However, since then, a number of systematic reviews of the international evidence *have* been undertaken, but to date they have not been systematically located, synthesised and quality appraised. This paper seeks to redress that gap in the literature. It outlines in detail the reading policy development, mainly in England, but with reference to international developments, in the last ten years. It then reports the design and results of a systematic 'tertiary' review of all the relevant systematic reviews and meta-analyses in order to provide the most up-to-date overview of the results and quality of the research on phonics.

Keywords: phonics; reading policy; systematic review

Introduction

Improving standards of literacy through education and schooling in particular is a shared objective for education globally. This is reflected in co-ordinated approaches to

measure improvement internationally such as through the Progress in International Reading Literacy Study (PIRLS) (Mullis *et al*., 2009). An increased policy focus on standards of literacy is also evident (e.g., Schwippert & Lenkeit, 2012), as well as on methods of initial teaching. In the initial teaching of reading in languages with highly consistent orthographies (e.g., Spanish and especially Finnish), phonics is used without comment or dispute as the obvious way to give children who are not yet reading the most effective method of 'word attack', identifying unfamiliar printed words. The teaching of early reading in English, by contrast, has been highly politicised and is contentious, largely because of its notoriously complex set of grapheme-phoneme correspondences. In the United States (US) the so-called 'reading wars' have seen phonics approaches set against whole language approaches in decades of debate. While there have been what might be called 'reading skirmishes' in the United Kingdom (UK), they do not seem to have reached the same level of acrimony.

In 2007, British government policy on how children should be taught to read changed. Until 2006, within the statutory National Curriculum (NC) for the teaching of English in state schools in England, the National Literacy Strategy recommended the so-called 'searchlights' model for teaching reading which was a 'mixed methods' approach, including embedded phonics, but also drawing on other approaches. From 2007 onwards, exclusive, intensive, systematic, explicit synthetic phonics instruction was adopted nationally. Also, and significantly, in 2007 this sentence: 'Children will be encouraged to use a range of strategies to make sense of what they read' was removed from the NC.

In 2006 two reviews on the teaching of reading funded by the Department for Education and Skills (D*f*ES) were published using alternative designs: a systematic review (SR) undertaken by two of the authors of this paper and a colleague (Torgerson

et al., 2006) and an expert review undertaken by Jim Rose (Rose, 2006). The SR used explicit transparent replicable methods, with systematic identification and inclusion of studies employing strong designs which can establish causal relationships between interventions and outcomes (randomised controlled trials or RCTs), minimisation of bias at every stage in the design and methods of the review, and assessment of the quality of the evidence base *before* coming to any conclusions. In contrast, the Rose Review did not use explicit methods for identification of studies to include and did not assess the quality of the evidence base, despite acknowledging the limitations of the UK-based trials (Rose, 2006, paragraphs 204 & 207, p. 61) included in his review.

In our systematic review, we found 12 individually randomised controlled trials; all were very small and only one was from the UK. In a meta-analysis, we found a small, statistically significant effect on reading accuracy, which we judged was derived from *moderate* weight of evidence, due to the relatively small number of trials and their variable quality. All the included studies integrated phonics with whole text level learning – in other words the phonics learning was not discrete. Our main recommendation was that systematic phonics instruction should be part of every literacy teacher's repertoire and a routine part of literacy teaching *in a judicious balance with other elements.* The difficulty of making policy recommendations for teaching reading is that such a 'judicious balance' may be disrupted by policy decisions that lack a reliable evidence base.

Background

The policy context: phonics in the National Curriculum for English in England

There have been three recognisable phases in the policy context in England since 1989. It should be noted that these apply only to England; Northern Ireland, Scotland and Wales have devolved responsibility for education.

Phase 1: Making phonics statutory

A National Curriculum (NC) for English in state schools in England was introduced in 1989, and there have been three subsequent versions (1995, 1999 and 2013). All covered the compulsory education years (ages 5 to 16), but only the sections for the primary years (ages 5 to 11) are relevant here. The first edition made just one reference to phonics: 'Pupils should be able to ... use picture and context cues, words recognised on sight and phonic cues in reading' (Department of Education and Science, 1989, p. 7). This appeared to place phonics on a par with other 'cue' systems for word recognition, even though those are little better than guessing since they often lead to learners producing words other than the target (see, in particular, Stanovich, 2000). Teaching children to rely on phonics to identify unfamiliar words would be more efficient.

Debate about the role and value of phonics was fuelled by the second (1989) edition of Chall's seminal *Learning to Read: The Great Debate* (1967), and by Adams' (1990) similarly comprehensive review; both concluded that phonics instruction enables children to make faster progress in (some aspects of) reading than no phonics or meaning-emphasis approaches, especially if applied to meaningful texts. Accordingly, the second edition of the NC (Department for Education, 1995, pp. 6-7) provided significantly more detail on phonics, while still giving a list of the 'key skills' for early reading that was essentially the same as in NC Mark 1. However, the essential terms for defining the process of phonics, namely 'phoneme' and 'grapheme', were not even

mentioned, let alone the necessary underpinnings in phonetics and analysis of grapheme-phoneme correspondences.

To support NC Mark 2, the National Literacy Strategy (NLS) was rolled out from 1997. The NLS *Framework for Teaching* (Department for Education and Employment (D*f*EE), 1998) at last introduced the term 'phoneme', but still portrayed phonics as just one of its 'searchlights' strategies for identifying words and comprehending text, the others being much the same as in NC Mark 1 and 2.

In the third edition of the NC (D*f*EE, 1999a, p. 46) the amount of detail on phonics was much the same as in the second edition, but more focused, including using 'phoneme'. Shortly afterwards, reports from the National Reading Panel (2000) and its phonics subgroup (Ehri *et al.*, 2001) appeared in the US, and slowly began to influence research and practice in Britain.

In its report on the first four years of the NLS, the Office for Standards in Education, Children's Services and Skills (Ofsted, 2002) praised some aspects of the teaching of phonics in primary schools in England but criticized others; even the fact that they could do this showed that there was more, and more focused, phonics teaching than a decade earlier. A set of support materials, *Playing with Sounds* (D*f*ES, 2004), was published soon afterwards. In a period of 15 years, therefore, phonics had moved from virtual invisibility to being a central concern, with statutory backing and professional guidance.

Phase 2: Which variety of phonics?

Johnston and Watson (2004) reported on two studies in Scotland comparing synthetic and analytic phonics. Experiment 1, which was not an RCT but a quasi-experiment, compared a synthetic phonics group with two analytic phonics groups and found an

advantage for the synthetic phonics group, *but* this group had received training at a faster pace than the others, and five of the 13 whole classes involved had been allocated by the researchers to receive synthetic phonics according to their perceived greater need.

Experiment 2, which was actually conducted *before* Experiment 1, also compared synthetic phonics and analytic phonics and found a positive effect for synthetic phonics, *but* one researcher taught both groups, and the researchers did not report their method of randomisation or their sample size calculation, did not undertake intention to treat analysis (the correct analysis, keeping children in their originally allocated groups), and did not use blinded assessment of outcome.

Despite these methodical flaws, publicity for Experiment 1 (Experiment 2 received very little) led many to believe that synthetic phonics had the edge, and attracted sufficient political attention for a parliamentary committee on to hold an enquiry into teaching children to read in 2004-05; its report (House of Commons Education and Skills Committee, 2005) appeared in the spring of 2005. In quick succession thereafter the British Government: commissioned the systematic review of the research evidence on phonics (Torgerson *et al*., 2006) which is the precursor of this 'tertiary' review; set up the Rose Review, which concentrated on good practice in the teaching of reading, including in the use of phonics, and reported in early 2006 (Rose, 2006); established a pilot project on synthetic phonics to begin in 2005; and commissioned the *Letters and Sounds* framework for phonics teaching which the D*f*ES itself published (D*f*ES, 2007).

In 2006 we built on the systematic review which had appeared in the US (Torgerson *et al*. 2006). Ehri *et al.* (2001; see especially p. 393) had analysed data from both RCTs and quasi-experiments; they concluded that systematic phonics instruction

enabled children to make better progress in reading than instruction featuring unsystematic or no phonics. However, they also concluded that there was no evidence to show that any particular form of phonics was superior to any other form of phonics. Using only RCTs, including the first from Britain (experiment 2 of Johnston and Watson, 2004), found firm evidence that systematic phonics instruction enables children to make better progress in *word recognition* than unsystematic or no phonics instruction, but not enough evidence to decide whether (a) systematic phonics instruction enables children to make better progress in *comprehension*, or (b) whether synthetic or analytic phonics is more effective (Johnston and Watson's experiment 2 was one of only three relevant RCTs).

Our first conclusion was welcome to the Rose committee, but not the second or particularly the third. However, Jim Rose and colleagues who made classroom observation visits in 2005 concluded that synthetic phonics is more effective. Rose's (2006) conclusion that systematic phonics equates with synthetic phonics was seized upon by opponents as going beyond the evidence – see, for example, the debate in *Literacy*, vol.41, no.3 (Brooks *et al.*, 2007). Though some opposition to phonics is still reported (e.g., most recently Krashen, 2017), some of it based on the misapprehension that there is a forced choice between phonics and whole-language approaches, that controversy seemed to die down within a few years, and the place of phonics as part of the initial teaching of literacy now seems largely accepted in England.

The rational way to investigate the relative effectiveness of synthetic and analytic phonics would have been to conduct a large and rigorous RCT (as advocated by us in 2006: see Torgerson *et al*., 2006:12). Instead, the pilot project on synthetic phonics alone, known as *The Early Reading Development Pilot*, began in the school year 2005/06 in 172 schools in 18 Local Authorities (LAs). Although no separate report on

that pilot seems ever to have been published, a decision was evidently taken in central government to roll synthetic phonics out nationally, and this was carried out in successive batches of LAs between 2006/07 and 2009/10, under the title *The Communication, Language and Literacy Development Programme*.

The results of these programmes seem to have been analysed and published only with the appearance of a report by Machin *et al.* (2016), who also had access to national pupil attainment data at ages 5, 7 and 11. By using the staggered roll-out to define quasi-'treatment' and 'control' groups, the authors were able to estimate the effect of introducing synthetic phonics on children's attainment at all three ages. They concluded that there had been an across-the-board improvement at ages 5 and 7, but that at age 11 there was no average effect – however, there were lasting effects for children who could be considered as having been at risk of underachievement initially (children who entered school at risk of falling behind, those who were from disadvantaged backgrounds, and non-native speakers of English – precisely the groups one would hope would benefit) (Machin *et al.*). This result means that there would have been a negative effect for the remaining children as there was no average overall effect.

The Rose report had contained a set of criteria for judging phonics teaching schemes, and in 2007-10 the D*f*ES supported two different panels providing quality assurance of publishers' claims about their schemes against those criteria (see Beard *et al.*, forthcoming); one of the mainly initial schemes judged was *Letters and Sounds*.

The Rose review also contained, in an appendix, a version of the 'Simple View of Reading' (Gough and Tunmer, 1986) by Morag Stuart, which she elaborated in Stuart (2006). This theory portrays reading comprehension as the product of language (listening) comprehension and the decoding of printed words, and holds that these dimensions can (largely) vary independently and that both decoding and comprehension

require explicit teaching. In the Primary National Strategy (D*f*ES, 2006), which had incorporated the NLS, this model of reading processes replaced the 'Searchlights' model.

So far, so largely similar, it would seem, to developments in other English-speaking countries. There was little remaining opposition to the use of phonics in initial literacy teaching, the Simple View of Reading had become the predominant model, and synthetic phonics had become the favoured variety, as later advocated and analysed in Stuart and Stainthorp (2016). But in England there was to be a significant further policy turn which does not seem to have been matched elsewhere and has caused renewed controversy.

Phase 3: Putting a strong official push behind synthetic phonics

There have been significant developments since the change of government in 2010. A third panel providing the DfE with quality assurance of publishers' claims about their phonics schemes operated in 2010-12; one of the criteria was re-worded to require that schemes be synthetic. Commercial publishers had to re-submit their schemes, and some which had passed the scrutiny of the earlier panels failed this time (see again Beard *et al.*, forthcoming). Almost half the roughly 100 schemes evaluated failed because they contained basic linguistic and/or phonetic errors (e.g. confusing graphemes and phonemes, or diphthongs and digraphs).

From September 2011 to October 2013, if schools ordered schemes which met the revised criteria and were therefore on an 'approved list' (in the form of a phonics catalogue on the DfE website), they could receive match funding from the DfE. In September 2014 there were just 10 full synthetic phonics schemes, and 15 sets of supplementary resources, on the DfE's approved list (DfE, 2014).

The most important development after the change of government was the introduction of the 'phonics screening check' for Year 1 pupils, which was piloted in the summer term 2011 and has been implemented nationally in each summer term since 2012 (for the background, see DfE, 2011). This individually-administered 'check', which is a test in all but name, was promoted as 'telling parents how well their children are getting on with learning to read', and consists of 40 letter-strings to be read aloud; half are real words, the rest non-words designed to assess whether children have mastered the grapheme-phoneme correspondences (GPCs) without which they would not be able to vocalise these items. Children who score below the 'threshold' or pass mark (32 correct out of 40) receive extra instruction during Year 2, and at the end of that year are re-tested; most pass on this second attempt, but some do not, and are not re-tested again in Year 3; nor is there (apparently) any further centrally-directed support for them. The test continues in force despite vocal opposition (e.g. Clark, 2015), and a detailed analysis (Darnell *et al.*, 2017) showing that some items require word knowledge in addition to ability to use GPCs, and that some GPCs listed in the government's specification are not in fact tested.

Meanwhile, a new version of the national curriculum was published in 2013 for implementation in 2014. It is worth quoting its two main statements on phonics:

> '[Year 1] Pupils should be taught to: apply phonic knowledge and skills as the route to decode words; respond speedily with the correct sound to graphemes (letters or groups of letters) for all 40+ phonemes, including, where applicable, alternative sounds for graphemes; read accurately by blending sounds in unfamiliar words containing GPCs [grapheme-phoneme correspondences] that have been taught…'

(DfE, 2013, p. 20)

> [Other relevant information includes:] ‘Skilled word reading involves both the speedy working out of the pronunciation of unfamiliar printed words (decoding) and the speedy recognition of familiar printed words. Underpinning both is the understanding that the letters on the page represent the sounds in spoken words. This is why phonics should be emphasised in the early teaching of reading to beginners (i.e. unskilled readers) when they start school.’
> (DfE, 2013, p. 4)

The first of these paragraphs contains a clear and distinctive summary of synthetic phonics for reading, and both paragraphs correctly define its use as being the identification of *unfamiliar* printed words. Taken with other statements in the curriculum concerning synthetic phonics for spelling (e.g., p. 29) and for reading in Year 2, the notion that phonics should effectively be complete by the end of Year 2, and the comprehension and enjoyment of reading, this is a balanced view. However, the curriculum also contains an appendix (pp. 49-73) laying out in great detail the principal phoneme-grapheme and grapheme-phoneme correspondences of British English spelling relative to the RP (Received Pronunciation) accent (with a few notes on regional variation, e.g. in the pronunciation of words like *bath* and *past*), and providing a key to the International Phonetic Alphabet symbols used (p. 73). While this knowledge appears essential for teachers to ensure accurate phonics teaching, the contrast with the exiguous earlier specifications of phonics is stark.

The overall picture of phonics in the National Curriculum for English in England is therefore of an initial tentative phase, followed by the deliberate choosing of synthetic

phonics before research evidence justified this, and now firm government pressure to ensure the implementation of that variety of phonics. How accurate that implementation is remains to be investigated, as does its continued effectiveness. The Machin *et al.* (2016) findings are based on data from 2004-11, and therefore pre-date both the Year 1 phonics test and NC Mark 4, with its highly detailed specifications. At the time of writing there is no sign that phase 3 has an end.

Rationale for the tertiary review

Ten years after the publication of our systematic review (Torgerson *et al*., 2006), the reading skirmishes are alive and well, and the UK-based RCT we recommended has never been undertaken. However, a number of SRs and meta-analyses (and methodological re-analyses of existing meta-analyses) *have* been undertaken since 2006, and a tertiary review is particularly helpful where a number of overlapping systematic reviews have been undertaken in a given topic area (as is the case with phonics) in order to explore consistency across the results from the individual reviews. A synthesis of the findings of these studies provides a more complete picture of the evidence for the effectiveness of phonics (or alternative) reading approaches in terms of a pooled effect size or narrative synthesis of quantified outcomes of the extant SRs, and is more robust than simply looking at individual systematic reviews, small scale RCTs or a *non*-systematic synthesis of previous SRs.

Design and methods

The most scientific approach to searching for, locating, quality appraising and synthesising all the relevant systematic reviews in a tertiary review is to use systematic review design and methods: an exhaustive and unbiased search; minimisation of bias at all stages of inclusion; data extraction and quality appraisal because this increases the overall reliability in the findings. We aimed to explore the consistency (or lack) of the findings across the full range of the located reviews. In addition, we wanted to look at methodological challenges with respect to: the quality of the reviews; publication bias; and the difference in results depending on both the designs and the statistical models used in the included studies.

We used SR methods at all stages of the tertiary review, including applying strict quality assurance procedures to ensure rigour and, consequently, to increase confidence in our results.

Primary research questions

What is the effectiveness of systematic phonics instruction compared with alternative approaches, including whole language approaches or different varieties of phonics on reading accuracy, comprehension and spelling; and what is the quality of the evidence base on which this judgement is formed?

Secondary research questions

Does the evidence for effectiveness vary by design and/or statistical model for effect size calculation? Is there evidence of publication bias in the included systematic reviews, and consequently in the tertiary review itself?

Inclusion/exclusion criteria

We established inclusion criteria prior to starting the search for studies. As a minimum, included SRs had to provide evidence of the three key items of a SR for an effectiveness question, namely: a systematic search primarily using electronic databases; quality appraisal of all included studies; and a quantified synthesis or meta-analysis giving pooled effect sizes. Systematic reviews also had to include studies using a rigorous design that is able to establish causal relationships between interventions and outcomes - experimental or quasi-experimental designs (RCTs and/or QEDs). In terms of interventions, we included reviews of studies evaluating the effectiveness of phonics interventions compared with whole-language interventions or alternative approaches, including different varieties of phonics instruction (synthetic or analytic). In terms of outcomes, we included reviews of studies that included any combination of any standardised reading and spelling outcomes.

Searching

The search strings were based on relevant key words and their derivatives. For example, in ASSIA, ERIC and PsycINFO they were as follows:

> (phonic* OR phonetical* OR phonemic) AND (systematic review OR meta-analysis OR research synthesis OR research review)

See Appendix A for the full search strategies for all databases searched in 2014 and 2016.

We searched exhaustively (from 2001) for all the potentially relevant systematic reviews, containing meta-analyses with pooled effect sizes. The databases searched were: Applied Social Sciences Index and Abstracts (ASSIA), Education Resources

Information Centre (ERIC), PsycINFO, Web of Science and World Cat. Searches were undertaken in 2014 and 2016.

Screening at first and second stages

We screened the titles and abstracts (first stage) and full papers (second stage) for inclusion using pre-established inclusion criteria. Independent double screening ensured a robust approach to this process.

Data extraction and quality appraisal

All included systematic reviews/meta-analyses were independently data-extracted and quality-appraised using specifically designed templates by two pairs of reviewers, who then conferred and agreed a final version. The template for data extraction included substantive items: details about the nature of included interventions and control conditions; number and designs of included studies; participants and settings; and outcome measures and results. The template for quality appraisal of included SRs included methodological items of the included SRs from the PRISMA checklist (Moher, Liberati, Tetzlaff and Altman, 2009), including: methods for each stage of the review, including assessment of risk of bias within and across studies. We also extracted onto specifically-designed templates data to enable us to investigate the potential for both publication bias and design bias.

Results

Results of searching

After de-duplication there were 369 hits for the 2014 searches and 83 hits for the 2016 update. In total we included 452 potentially relevant studies from the electronic searching. Table 1 and the PRISMA diagram in Appendix A show the results from searching all the databases at the two time points.

[INSERT TABLE 1 HERE]

Results of screening

After screening of titles and abstracts and full papers we included a total of 12 studies. Table 2 and the completed PRISMA diagram in Appendix A show the results from screening at both stages. We found a total of 12 studies that met our inclusion criteria for the period 2001 to 2016.

[INSERT TABLE 2 HERE]

Results of quality assurance of screening

Initial agreement between the two authors who screened the entire database was high at both first and second stages. Any disagreements were resolved through discussion.

[INSERT TABLE 3 HERE]

Results: Characteristics and quality of SRs/meta-analyses

In Table 3 we summarise the main characteristics of the 12 SRs. Half (6) were undertaken in the United States, with one each in the United Kingdom and Australia, three in Germany, and one jointly in the US and Canada. Although many of the SRs

focused solely on the effectiveness of phonics interventions compared with control or comparison conditions, a number looked more broadly at a range of strategies to improve reading and spelling, with phonics instruction as a sub-category (see Table 3 for specific phonics interventions).

Most of the studies provided enough detail of the interventions included to show that almost all of those labelled 'phonics' were indeed investigating approaches to the teaching of reading and spelling which focus on letter-sound relationships, i.e. the association of phonemes with graphemes. However, Adesope *et al.* (2011) were vague on this point, and McArthur *et al.* (2012) used such a narrow definition of 'pure' phonics that only three studies qualified. Galuschka *et al.* (2014) and Han (2009) included pedagogies which would not qualify as phonics by any reasonable professional definition – it is therefore questionable whether they should have been included in this review. Other authors may also have included non-phonics studies, but it was beyond the scope of this review to check back to every individual RCT.

A few authors (Han, 2009; McArthur *et al*., 2012; Suggate, 2010, 2016) compared phonics instruction with phonemic/phonological awareness training. Details of the instruction received by control groups were scant; where mentioned, it seemed to be 'business as usual' literacy teaching, often of a whole language variety, though McArthur *et al*. (2012) and Suggate (2010) hinted at alternative interventions (e.g., maths).

The number of studies included in the SRs ranged from 3 to 85, so the various SR authors were clearly using different definitions of phonics and/or inclusion/exclusion criteria. Some of the variation was due to participant selection – e.g., Adesope *et al.* (2011) were looking at ESL students in English-speaking countries. Only Galuschka *et*

al. (2014) and Suggate (2010) included studies conducted in languages other than English. Participants in the studies included in the SRs range in age from pre-kindergarten children (aged 4), through children in all grades in primary (and middle) and secondary (high) schools, to adult participants in one SR. The full range of learner characteristics is represented in one or more SRs, including normally attaining and low-attaining students, those with English as a second language, or those with reading disabilities. Outcome measures in the SRs were diverse but most included studies with reading (decoding, word reading and fluency; comprehension) and spelling (writing).

[INSERT TABLE 4 HERE]

Table 4 presents the results of our quality assessment of the included SRs, using the key methodological items from the PRISMA statement. The 12 SRs were of generally high, but variable quality. Most of the 12 SRs fulfilled the following criteria by providing data or text: the rationale and objectives of the SR; methods and results for searching, screening, data collection and synthesis. (The three replication SRs used the databases from the original SRs for inclusion). Having said that, a key item from the PRISMA checklist – assessment of risk of bias of included studies – was undertaken by only 7 out of the 12 SRs. In other words, 5 of the SRs did not quality appraise the studies which they included in their systematic review – and by extension, their pooled effect size – so they may have been indiscriminately including studies of high, moderate and low quality. This omission in these 5 SRs is critical and, therefore, the results from these SRs should carry lower weight of evidence in our conclusions.

[INSERT TABLE 5 HERE]

Results of effect sizes for phonics

Statistically significant positive effects for phonics instruction on at least one reading outcome were found across most (10) of the SRs ranging from small to moderate effects (Adesope *et al.*, 2011; Camilli *et al.*, 2003; Ehri *et al.*, 2001; Galuschka *et al.*, 2014; Han, 2009; McArthur et al, 2012; Sherman, 2007; Suggate, 2010; Suggate, 2016; Torgerson *et al.*, 2006). Non-significant positive effects were found in the remaining 2 SRs (Camilli *et al.*, 2006; Hammill and Swanson, 2006).

Effect size variance according to statistical model – Hedges' g or Cohen's d

The extracted effect sizes were classified according to how they were described by the authors. Most studies described or referenced the formulae for the effect size calculations and referred to this as *g* (Adesope *et al.*, 2011; Galuschka *et al.*, 2014; Han, 2009) or *d* (Ehri *et al.*, 2001, by cross-reference to NRP, 2000 – see footnote to Table 5); McArthur *et al.*, 2012; Sherman, 2007; Torgerson *et al*., 2006). One author (Suggate, 2010, 2016) followed Hunter and Schmidt's (2004) approach. Three studies used or referred to the approach adopted in the studies they were critiquing or defending (Camilli *et al.*, 2003, 2006; Hammill & Swanson, 2006).

There is some confusion in the literature about terminology, but Hedges' *g* usually refers to Hedges' bias-corrected estimator (Hedges and Olkin, 1985) and *d* to Cohen's *d* (Cohen, 1988). Both approaches are based on a pooled standard deviation. Cohen used the maximum likelihood estimator for the variance, which is biased with small samples, whereas Hedges used Bessel's correction (n-1) to estimate the variance. In practice, for samples above 20 the difference in the effect size estimate is minimal. Estimates of effect will also vary between class and individual level analysis,

and depending on whether unequal sample sizes and clustering are taken into account (Xiao, Kasim & Higgins, 2016), and on which mean scores are used (post-test or gains) and on which standard deviations are pooled (pre-test, post-test or gains). Some further details can be found in Table 5.

However, it should be noted that, of all the SRs reviewed, only Galuschka *et al.* (2014, p. 3) stated which mean scores were used in calculating ESs (post-test); they implied that the pooled standard deviations used were those of the post-test. The hidden problem when authors do not report these details is that even various results labelled as 'Cohen's d' or 'Hedges' g' may not be strictly commensurate with each other, and this may bedevil attempts to generalise from them.

Effect size variance according to design – RCT or QED

The included SRs contained both RCTs and QEDs, with two exceptions (Galuschka *et al.*, 2014; Torgerson *et al*., 2006) which included only RCTs. In two cases it was not possible to determine which studies were of which designs (Adesope, 2011; Sherman, 2007). In a number of the included SRs the authors did not report study design for the studies which investigated the effectiveness of phonics instruction. Looking at the pooled effect sizes (ES) from RCTs and QEDs, for those reviews that have included both, there are some clear differences. Some of these differences in ES are less apparent in the overall reported ES. For example, as Table 5 shows, Adesope *et al.* (2011) do not explicitly report ES separately for RCTs and QEDs; however, the pooled ES for random allocation is +0.31 and +0.68 for non-random allocation, a difference of +0.37. This difference is less apparent in looking at the pooled overall ESs; that for systematic phonics instruction and guided reading is +0.40 and that collapsed across all pedagogical strategies is +0.41. Suggate (2010) is similar, in that the overall ES for

QEDs is larger (+0.64) than for RCTs (+0.41), with the overall mean weighted ES for phonics being +0.50. Camilli *et al.* (2003) explicitly stated that there was no difference between ES for RCT and QED designs, with an overall ES of +0.24. Similarly, different ES are not stated in Camilli *et al.* (2006) for different designs; the overall ES reported is, however, much lower at +0.12.

Publication bias

We extracted data from each study about whether or not grey literature was searched; whether any grey literature was included; whether the issue of publication bias seemed to have the potential to bias the results of the study; whether a recognised method for the detection of publication bias was used (for example, funnel plot); whether any evidence for potential publication bias was found; and, if publication bias was suspected, what method was used to mitigate this bias and the results flowing from this (see Table 6).

Of the 12 systematic reviews, only 6 engaged fully with the issue of publication bias and the potential for it to bias the results of their systematic review (Adesope *et al.*, 2009; Galuschka *et al.*, 2014; McArthur, 2012; Suggate, 2010, 2016; Torgerson *et al*., 2006). The remaining 10 studies either did not mention publication bias at all (or this was unclear) or, as in the case of Han (2009), publication bias was mentioned but the author did not search for or include any grey literature, and did not use any method to assess the potential for publication bias. Sherman *et al.* (2007) searched for grey literature, but had as an exclusion criterion 'not published in peer-reviewed journals' and therefore excluded those studies that they had retrieved but which were not published (total of 5). They also did not mention the issue of publication bias, in

particular that the application of the exclusion criterion may have contributed to publication bias in their review.

Adesope *et al.* (2009) did not search for or include any grey literature. However, they did explore the issue through the use of Orwin's Fail-Safe N and Classic fail-safe N test, which suggested that the results were robust and validity was not threatened by publication bias; therefore no further analyses were undertaken.

Galuschka *et al.* (2014) explored publication bias for those studies which evaluated phonics instruction and used reading performance as a dependent variable (not for spelling). A funnel plot was used to explore the presence of publication bias, which displayed asymmetry with a gap on the left of the graph, indicating the possible presence of publication bias. Duval and Tweedie's trim and fill method was used to assess the extent of publication bias, and an unbiased effect size was estimated. The procedure trimmed 10 studies into the plot and led to an estimated unbiased effect size of Hedges' g = +0.198 (CI +0.039, +0.357), which is in contrast to a, potentially biased upwards, effect size of Hedges' g = +0.32 (CI +0.18, +0.47) for the main analysis.

McArthur *et al.* (2012) searched for and included grey literature and also undertook sensitivity analysis and a funnel plot, and concluded that their systematic review was not affected by publication bias.

Although he did not explicitly search for and include studies from the grey literature, in two meta-analyses Suggate (2010, 2016) looked at the potential for publication bias using funnel and box plots, and addressed this in the more recent meta-analysis by including only the larger studies.

In our SR (Torgerson *et al*., 2006) we specifically searched the grey literature, and included one unpublished thesis. They used a funnel plot to investigate the potential

presence of publication bias in their meta-analysis and found evidence of this, but the Egger test statistic was not significant, which reduced any certainty in their finding.

[INSERT TABLE 6 HERE]

Results of quality assurance of data extraction and quality appraisal

Initial agreement between the two pairs of authors was high; any disagreements were resolved through discussion and arbitration. The data extraction and quality appraisal of the original SR undertaken by two of the authors Torgerson *et al*. (2006) were completed by the other two authors to minimise the potential for conflict of interest.

Discussion

The diverse range of interventions and control or comparison conditions, settings (including countries), participant characteristics, outcome measures and study designs included in the 12 SRs in our tertiary review increases the generalizability of our findings. However, there are limitations on this, in particular doubts over whether some of the interventions analysed deserve the label 'phonics', and the possible incommensurability of the overall effect sizes reported due to both under-reporting of, and differences in, methods of calculating them.

In terms of publication bias, as only 6 of the 12 meta-analyses addressed this issue and, of those, only 3 found evidence of potential publication bias, we can interpret this as an indication that publication bias is an issue in the individual meta-analyses in the tertiary review, and therefore in the tertiary review itself. The consequences of this interpretation are that we should have more caution in the findings of our review as it is

likely that experimental studies have been undertaken which have found null or negative results and therefore have either not been published, or they have been published but have not been included in meta-analyses, either by design or because they were not in the public domain to be found.

The reviews were fairly consistent in demonstrating an overall positive effect of phonics teaching, with pooled estimates ranging from 0.12 to 0.5. This is probably unsurprising, given that the reviews contained many of the same studies and therefore it would be unlikely that there would be huge divergence in terms of the pooled estimate. Furthermore, there is little evidence to demonstrate the superiority of one phonics approach compared with any other instructional method – but very few individual RCTs have investigated this question, so it hardly features in the SRs. There remains uncertainty as to the overall effect given the probable presence of publication bias. Indeed, with the prevalence of so many reviews showing positive effects of phonics teaching, this means it might be less likely for null or negative results to be reported.

Some of the reviews try to distinguish differential effects of phonics among educationally important subgroups. Whilst some reviews see some evidence for better or lesser effects within different types of learner, these forms of analysis should always be treated with a certain amount of caution. This is because even, within a large randomised controlled trial, there is usually very little statistical power to demonstrate meaningful subgroup differences, and within a meta-analysis the power issue is even more problematic.

Conclusions

Given the evidence from this tertiary review, what are the implications for teaching, policy and research? It would seem sensible for teaching to include systematic phonics instruction for younger readers – but the evidence is not clear enough to decide which phonics approach is best. Also, in our view there remains insufficient evidence to justify a 'phonics only' teaching policy; indeed, since many studies have *added* phonics to whole language approaches, balanced instruction is indicated. For policy, encouragement of phonics instruction within schools is justified unless and until contrary evidence emerges. Finally, in terms of research: given the uncertainties in the evidence base over publication bias, the 'phonics' status of some included studies, and how best to calculate effect sizes, there may be a case for conducting a large and even more rigorous systematic review. But what is required above all are large field trials of different phonics approaches and different phonics 'dosages'. We called for such an approach in our review of phonics teaching in 2006, and a decade later we make the same call.

In conclusion, there have been a significant number of systematic reviews of experimental and quasi-experimental research evaluating the effectiveness or otherwise of phonics teaching since 2000. Most of the reviews are supportive of phonics teaching, but this conclusion needs to be tempered by two potential sources of bias: design and publication bias. Both of these problems will tend to exaggerate the benefit of phonics teaching. Furthermore, there is little evidence of the comparative superiority of one phonics approach over any other. Ideally, each country should establish a programme of large RCTs that are adapted to local circumstances that will test different phonics approach to reading and writing acquisition. If this was adopted then we might finally end the 'reading wars'.

Included systematic reviews/meta-analyses

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2011). Pedagogical strategies for teaching literacy to ESL immigrant students: A meta-analysis. *British Journal of Educational Psychology,* 81(4), 629-653.

Camilli, G., Vargas, S., & Yurecko, M. (2003). "Teaching Children To Read": The Fragile Link between Science and Federal Education Policy. *Education Policy Analysis Archives,* 11(15).

Camilli, G., Wolfe, P. M., & Smith, M. L. (2006). Meta-Analysis and Reading Policy: Perspectives on Teaching Children to Read. *Elementary School Journal, 107*(1), 27-37.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic Phonics Instruction Helps Students Learn to Read: Evidence from the National Reading Panel's Meta-Analysis. *Review of Educational Research,* 71(3), 393-447.

Galuschka, K., Ise, E., Krick, K., & Schulte-Koerne, G. (2014). Effectiveness of Treatment Approaches for Children and Adolescents with Reading Disabilities: A Meta-Analysis of Randomized Controlled Trials. *Plos One,* 9(2).

Hammill, D. D., & Swanson, L. H. (2006). The National Reading Panel's Meta-Analysis of Phonics Instruction: Another Point of View. *Elementary School Journal,* 107(1), 17-27.

Han, I. (2010). *Evidence-based reading instruction for English language learners in preschool through sixth grades: A meta-analysis of group design studies.* University of Minnesota, ProQuest Dissertations Publishing, 2009. 3371852

McArthur, G., Eve, P. M., Jones, K., Banales, E., Kohnen, S., Anandakumar, T., . . . Castles, A. (2012). Phonics training for English-speaking poor readers. *Cochrane Database of Systematic Reviews* (12 December 2012).

Sherman, K. H. (2007). A meta-analysis of interventions for phonemic awareness and phonics instruction for delayed older readers. University of Oregon, ProQuest Dissertations Publishing, 2007. 3285626.

Suggate, S. P. (2010). Why What We Teach Depends on When: Grade and Reading Intervention Modality Moderate Effect Size. *Developmental Psychology,* 46(6), 1556-1579.

Suggate, S. P. (2016). A Meta-Analysis of the Long-Term Effects of Phonemic Awareness, Phonics, Fluency, and Reading Comprehension Interventions. *Journal of Learning Disabilities,* 49(1), 77-96.

Torgerson, C., Brooks, G., & Hall, J. (2006). A systematic review of the research literature on the use of phonics in the teaching of reading and spelling (ISBN: 1844786595 9781844786596). Retrieved from http://catalogue.bishopg.ac.uk/custom_bgc/files/JKEC_phonics_review.pdf

References

Adams, M.J. (1990). *Beginning to Read: Thinking and Learning about Print*. Cambridge, MA: MIT Press.

Beard, R., Brooks, G. & Ampaw-Farr, J. (forthcoming). How linguistically-informed are phonics programmes?
Literacy.

Brooks, G., Cook, M. & Littlefair, A., with replies from Wyse, D. & Styles, M. (2007). Responses to Wyse and Styles' article, "Synthetic phonics and the teaching of reading: the debate surrounding England's 'Rose Report' " (*Literacy*, 41, 1, April 2007). *Literacy*, 41, 3, 169-76.

Chall, J.S. (1967) *Learning to Read: The Great Debate.* New York, NY: McGraw-Hill. Second edition, 1989.

Clark, M. (2015). An evidence-based critique of synthetic phonics in literacy learning. *Primary First*, issue 12 (Spring).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Darnell, C.A., Solity, J.E. & Wall, H. (2017). Decoding the phonics screening check. *British Educational Research Journal*, 43, 3, 505-27.

Department of Education and Science (1989). *English in the National Curriculum.* London: Her Majesty's Stationery Office.

Department for Education (1995). *English in the National Curriculum.* London: Her Majesty's Stationery Office.

DfE (2011). *Year 1 phonics screening check pilot evaluation*. London: Department for Education. https://www.gov.uk/government/publications/year-1-phonics-screening-check-pilot-evaluation (accessed 5 February 2017)

DfE (2013). *English programmes of study: key stages 1 and 2 National curriculum in England*. London: Department for Education.

DfE (2014). *Phonics: Choosing a Programme*. London: Department for Education. https://www.gov.uk/government/collections/phonics-choosing-a-programme (accessed 5 February 2017)

D*f*EE (1998). *National Literacy Strategy*. London: Department for Education and Employment.

D*f*EE (1999). *The National Curriculum Handbook for primary teachers in England.* London: Department for Education and Employment & Qualifications and Curriculum Authority.

D*f*ES (2004). *Playing with Sounds*. London: Department for Education and Skills.

D*f*ES (2006). *Primary National Strategy*. London: Department for Education and Skills.

D*f*ES (2007). *Letters and Sounds*. London: Department for Education and Skills.

Ehri, L.C., Nunes, S.R., Stahl, S.A. & Willows, D.M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis.
Review of Educational Research, 71, 3, 393-447.

Gough, P. & Tunmer, W. (1986). Decoding, reading, and reading disability.
Remedial and Special Education, 7, 6–10.

Hartung, J., Knapp, G. & Sinha, G.M. (2008). *Statistical Meta-Analysis with Application*. Hoboken, New Jersey: Wiley.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis.*
New York, NY: Academic Press.

House of Commons Education and Skills Committee (2005). *Teaching children to read (Eighth report of Session 2004–05).* London: The Stationery Office Limited.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, Ca: Sage Publications.

Johnston, R.S. & Watson, J.E. (2004). Accelerating the development of reading, spelling and phonemic awareness skills in initial readers. *Reading and Writing*, 17, 327.

Krashen, K. (2017, 2 February). Letter in *The Guardian* newspaper. https://www.theguardian.com/education/2017/feb/01/invest-in-libraries-not-phonics-tests

Machin, S. McNally, S. & Viarengo, M. (2016). *"Teaching to Teach" Literacy*. London: London School of Economics Centre for Economic Performance Discussion Paper No 1425.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med*, *6*(7), e1000097. doi:doi:10.1371/journal.pmed.1000097

Mullis, I.V., Martin, M.O., Kennedy, A.M., Trong, A.L.& Sainsbury, M. (2009) *PIRLS 2011 Assessment Framework:* International Association for the Evaluation of Educational Achievement.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for*

reading instruction. Washington DC: National Institute for Child Health and Human Development Clearinghouse.

Ofsted (2002). *The National Literacy Strategy: the first four years 1998-2002*. London: Office for Standards in Education.

Rose, J. (2006). *Independent Review of the Teaching of Early Reading. Final report*. London: Department for Education and Skills.

Schwippert, K. & Lenkeit, J. (Eds.) (2012) *Progress in reading literacy in national and international context. The impact of PIRLS 2006 in 12 countries.* Munster: Waxmann Verlag.

Stanovich, K.E. (1988). Explaining the differences between the dyslexic and the garden-variety poor reader: the phonological-core variable-difference model. *Journal of Learning Disabilities*, 21, 10, 590-604.

Stanovich, K. E. (2000). *Progress in Understanding Reading: Scientific Foundations and New Frontiers*. New York: Guilford Press.

Stuart, M. (2006). Teaching reading: Why start with systematic phonics teaching? *Psychology of Education Review,* 30, 6-17.

Stuart, M. & Stainthorp, R. (2016). *Reading Development & Teaching*. London: Sage.

Xiao, Z., Kasim, A. & Higgins, S. (2016). Same difference? Understanding variation in the estimation of effect sizes from educational trials.

International Journal of Educational Research, 77, 1-14.

[INSERT APPENDIX A HERE]